IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS : Toby Walker et al.
SERIAL No. : 09/647,301 Group Art Unit : 2616
FILING DATE : November 21, 2000 Examiner : TRAN, THAI Q
TITLE : SIGNAL PROCESSING METHOD AND VIDEO SIGNAL PROCESSOR

Hon. Commissioner of Patents and Trademarks,
Washington, D.C. 20231

SIR:

CERTIFIED TRANSLATION

I, Takashi Narita, am an official translator of the Japanese language into the English language and I hereby certify that the attached comprises an accurate translation into English of Japanese Application No. 11-023064, filed on January 29, 1999.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

December 24, 2004 Takashi Narita

Date Takashi Narita

Patent Office

Japanese Government

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: January 29, 1999

Application Number: Patent Application
Ser. No. 11-023064

Applicant: Sony Corporation

November 19, 1999

Commissioner,

Patent Office Takahiko Kondo

[Document Name] Patent Application

[Reference Number] 9801128002

[To] Hon.Commissioner, Patent Office

[IPC] G01B 11/24

[Inventor]

[Address] c/o Sony Corporation
7-35, Kitashinagawa 6-chome,
Shinagawa-ku, Tokyo, Japan

[Name] Toby Walker

[Inventor]

[Address] c/o Sony Corporation
7-35, Kitashinagawa 6-chome,
Shinagawa-ku, Tokyo, Japan

[Name] Hiroshi Matsubara

[Patent Applicant]

[Identification Number] 000002185

[Name] Sony Corporation

[Representative] Nobuyuki Idei

[Patent Attorney]

[Identification Number] 100067736

[Patent Attorney]

[Name] Akira Koike

[Patent Attorney]

[Identification Number] 100086335

[Patent Attorney]

[Name] Eiichi Tamura

[Patent Attorney]

[Identification Number] 100096677

[Patent Attorney]

[Name] Seiji Iga

[Indication of Charge]

[Number of Prepaid Ledger] 019530

[Amount] 21,000 yen

[List of Document]

[Document] Specification 1

[Document] Drawing 1

[Document] Summary 1

[General Power of Attorney Number] 9707387

[Need of Proof] Yes

[Name of Document] SPECIFICATION

[Title of the Invention]

Signal Processing Method, and Video Signal Processor

[Claims]

[Claim 1]

A signal processing method for detecting and analyzing a pattern reflecting the semantics of the content of a signal, the method comprising steps of:

extracting, from a segment consisting of a sequence of consecutive frames forming together the signal, at least one feature information which characterizes the properties of the segment;

calculating, using the extracted feature information, a criterion for measurement of a similarity between a pair of segments for every extracted feature information and measuring a similarity between a pair of segments according to the similarity measurement criterion; and

detecting, according to the feature information and similarity measurement criterion, two of the segments, whose mutual time gap is within a predetermined temporal threshold and mutual similarity is not less than a predetermined dissimilarity threshold, and grouping the segments into a scene consisting of a sequence of temporally consecutive segments reflecting the semantics of the signal content.

[Claim 2]

The method as set forth in Claim 1, wherein the signal is at least one of visual

and audio signals included in a video data.

[Claim 3]

The method as set forth in Claim 1, wherein at the feature extracting step, a single statistic central value of the feature information at different time points in a single segment which is divided plurally is selected for extraction.

[Claim 4]

The method as set forth in Claim 1, wherein a statistic value of the similarity between a plurality of segment pairs is used to determine the dissimilarity threshold.

[Claim 5]

The method as set forth in Claim 1, wherein of the segments, more than at least one segment which could not have been grouped into a scene at the grouping step are grouped into a single scene.

[Claim 6]

The method as set forth in Claim 1, wherein a scene from arbitrary feature information acquired at the grouping step and more than at least one scene for feature information different from the arbitrary feature information, are combined together.

[Claim 7]

The method as set forth in Claim 2, wherein more than at least one scene from the video signal acquired at the grouping step and more than at least one scene from the audio signal acquired at the grouping step, are combined together.

[Claim 8]

A video signal processor for detecting and analyzing a visual and/or audio pattern reflecting the semantics of the content of a supplied video signal, the apparatus comprising:

means for extracting, from a visual and/or audio segment consisting of a sequence of consecutive visual and/or audio frames forming together the video signal, at least one feature information which characterizes the properties of the visual and/or audio segment;

means for calculating, using the extracted feature information, a criterion for measurement of a similarity between a pair of visual segments and/or audio segments for every extracted feature information and measuring a similarity between a pair of visual segments and/or audio segments according to the similarity measurement criterion; and

means for detecting, according to the feature information and similarity measurement criterion, two of the visual segments and/or audio segments, whose mutual time gap is within a predetermined temporal threshold and mutual similarity is not less than a predetermined dissimilarity threshold, and grouping the visual segments and/or audio segments into a scene consisting of a sequence of temporally consecutive visual segments and/or audio segments reflecting the semantics of the video signal content.

[Claim 9]

The apparatus as set forth in Claim 8, wherein the feature extracting means selects, for extraction, a single statistic central value of the feature information at different time points in a single visual and/or audio segment which is divided plurally.

[Claim 10]

The apparatus as set forth in Claim 8, wherein a statistic value of the similarity between a plurality of visual and/or audio segment pairs is used to determine the dissimilarity threshold.

[Claim 11]

The apparatus as set forth in Claim 8, wherein of the visual and/or audio segments, more than at least one visual and/or audio segment which could not have been grouped into a scene by the grouping means are grouped into a single scene.

[Claim 12]

The apparatus as set forth in Claim 8, wherein a scene for arbitrary feature information acquired by the grouping means and more than at least one scene for feature information different from the arbitrary feature information, are combined together.

[Claim 13]

The apparatus as set forth in Claim 8, wherein more than at least one scene from the visual signal of the video signal acquired by the grouping means and more than at least one scene from the audio signal of the video signal acquired by the grouping means, are combined together.

[Detailed Description of the Invention]

[0001]

[Technical Field of the Invention]

The present invention relates to a signal processing method for detecting and analyzing a pattern reflecting a semantics on which a signal is based, and a video signal processor for detecting and analyzing a visual and/or audio pattern reflecting a semantics on which a video signal is based.

[0002]

[Prior Art]

It is often desired to search, for playback, a desired part of a video application composed of a large amount of different video data, such as a television program recorded in a video recorder, for example.

[0003]

As a typical one of the image extraction techniques to extract a desired visual content, there has been proposed a story board which is a panel formed from a sequence of images defining a main scene in a video application. Namely, a story board is prepared by decomposing a video data into so-called shots and displaying representative images of the respective shots. Most of the image extraction techniques are to automatically detect and extract shots from a video structure as disclosed in "G. Ahanger and T. D. C. Little: A Survey of Technologies for Parsing and Indexing Digital Video, Journal of Visual Communication and Image Representation 7: 28-4,

1996", for example.

[0004]

[Problem to be Solved by the Invention]

It should be noted that a typical half-hour TV program for example contains hundreds of shots. Therefore, with the above conventional image extraction technique of G. Ahanger and T. D. C. Little, the user has to examine a story board having listed therein enormous shots having been extracted. Understanding of such a story board will be a great burden to the user. Also, a dialogue scene in which for example two persons are talking will be considered here. In the dialogue, the two persons are alternately shot by a camera each time either of them speaks. Therefore, many of such shots extracted by the conventional image extraction technique are redundant. The shots contain many useless information since they are at too low level as objects which are extracted from a video structure. Thus, the conventional image extraction technique cannot be said to be convenient for extraction of such shots by the user.

[0005]

In addition to the above, further image extraction techniques have been proposed as disclosed in "A. Merlino, D. Morey and M. Maybury: Broadcast News Navigation Using Story Segmentation , Proceeding of ACM Multimedia 97, 1997" and the Japanese Unexamined Patent Publication No. 10-136297, for example. However, these techniques can only be used with very professional knowledge of limited genres of contents such as news and football game. These conventional image

extraction techniques can assure a good result when directed for such limited genres but will be of no use for other than the limited genres. Such limitation of the techniques to special genres makes it difficult for the technique to easily prevail widely.

[0006]

Further, there has been proposed a still another image extraction technique as disclosed in the United States Patent No. 5,708,767 for example. It is to extract a so-called story unit. However, this conventional image extraction technique is not any completely automated one and thus a user's intervention is required to determine which shots have the same content. Also this technique needs a complicated computation for signal processing and is only applicable to video information.

[0007]

Furthermore, a still another image extraction technique has been proposed as in the Japanese Unexamined Patent Publication No. 9-214879, for example, in which shots are identified by a combination of shot detection and silent period detection. However, this conventional technique identifies a pair of shots between two silent periods and it can be used only when the silent period corresponds with a boundary between shots.

[0008]

Moreover, a yet another image extraction technique has been proposed as disclosed in "H. Aoki, S. Shimotsuji and O. Hori: A Shot Classification Method to

Select Effective Key-Frames for Video Browsing, IPSJ Human Interface SIG Notes, 7:43-50, 1996" and the Japanese Unexamined Patent Publication No. 9-93588 for example, in which repeated similar shots are detected to reduce the redundancy of the depiction in a story board. However, this conventional image extraction technique is only applicable to visual information, not to audio information.

[0009]

Accordingly, the present invention has an object to overcome the above-mentioned drawbacks of the prior art by providing a signal processing method and video signal processor, which can extract a high-level video structure in a variety of video data.

[0010]

[Means to Solve the Problem]

The above object can be attained by providing a signal processing method for detecting and analyzing a pattern reflecting the semantics of the content of a signal, the method including, according to the present invention, steps of: extracting, from a segment consisting of a sequence of consecutive frames forming together the signal, at least one feature information which characterizes the properties of the segment; calculating, using the extracted feature information, a criterion for measurement of a similarity between a pair of segments for every extracted feature information and measuring a similarity between a pair of segments according to the similarity measurement criterion; and detecting, according to the feature information and

similarity measurement criterion, two of the segments, whose mutual time gap is within a predetermined temporal threshold and mutual similarity is not less than a predetermined dissimilarity threshold, and grouping the segments into a scene consisting of a sequence of temporally consecutive segments reflecting the semantics of the signal content.

[0011]

In the above signal processing method according to the present invention, similar segments in the signal are detected and grouped into a scene.

[0012]

Also the above object can be attained by providing a video signal processor for detecting and analyzing a visual and/or audio pattern reflecting the semantics of the content of a supplied video signal, the apparatus including according to the present invention: means for extracting, from a visual and/or audio segment consisting of a sequence of consecutive visual and/or audio frames forming together the video signal, at least one feature information which characterizes the properties of the visual and/or audio segment; means for calculating, using the extracted feature information, a criterion for measurement of a similarity between a pair of visual segments and/or audio segments for every extracted feature information and measuring a similarity between a pair of visual segments and/or audio segments according to the similarity measurement criterion; and means for detecting, according to the feature information and similarity measurement criterion, two of the visual segments and/or audio

segments, whose mutual time gap is within a predetermined temporal threshold and mutual similarity is not less than a predetermined dissimilarity threshold, and grouping the visual segments and/or audio segments into a scene consisting of a sequence of temporally consecutive visual segments and/or audio segments reflecting the semantics of the video signal content.

[0013]

In the above video signal processor according to the present invention, similar visual segments and/or audio segments in the video signal are detected and grouped for output as a scene.

[0014]

[Embodiment of the Invention]

The embodiment of the present invention will further be described below with reference to the accompanying drawings.

[0015]

The embodiment of the present invention is a video signal processor in which a desired content is automatically detected and extracted from a recorded video data. Before going to the further description of the video signal processor, a video data to which the present invention is applicable will first be described.

[0016]

FIG. 1 shows a video data model having a hierarchy having three levels such as frames, segments and scenes, to which the present invention is applicable. As seen, the

video data model includes a sequence of frames at the lowest level. Also the video data model includes a sequence of consecutive segments at a level one step higher than the level of the frames. Further, the video data model includes scenes at the highest level. That is, a video data is formed from the scenes each consisting of the segments grouped together based on a meaningful relation between them.

[0017]

The video data includes both visual information and audio information. That is, the frames in the video data include visual frames each being a single still image, and audio frames representing audio information having generally been sampled for a time as short as tens to hundreds of milliseconds.

[0018]

As in the video data model, each of segments is comprised of a sequence of visual frames having consecutively been picked up by a single camera and these segments include visual segments and/or audio segments generally called "shots". Each of the segments is the fundamental unit of a video structure. Especially, the audio segments among these segments can be defined in many different manners as will be described below by way of example. For example, audio segments are bounded by periods of silence, respectively, in a video data detected by the well-known method, as the case may be. Also, in some cases, each audio segment is formed from a sequence of audio frames classified in several categories such as speech, music, noise, silence, etc. as disclosed in "D. Kimber and L. Wilcox: Acoustic Segmentation for

Audio Browsers, Xerox Parc Technical Report". Further, in other cases, the audio segments are determined using an audio cut detection to detect a large variation of a certain feature from one to the other of two successive audio frames, as disclosed in " S. Pfeiffer, S. Fischer and E. Wolfgang: Automatic Audio Content Analysis, Proceeding of ACM Multimedia 96, Nov. 1996, pp21-30", for example.

[0019]

Further, to group the content of a video data at a higher level including its semantics, the scene is made up of a meaningful group of features as feature information of segments such as perceptual activities in the segments which acquired visual frames and audio frames by detecting visual segments or shots or audio segments. The scene is subjective and depends upon the genre of content of the video data. The scene referred to herein is a group of repeated patterns of visual or audio segments whose features having some visual or audio relationships are similar to each other. More specifically, in a scene of a dialogue between two persons for example, visual segments appear alternately each time one of them speaks as shown in FIG. 2. In a video data having such a repeated pattern, a sequence of visual segments A of one of the two speakers and a sequence of visual segments B of the other speaker are grouped into one scene. The repeated pattern has a close relation with a hihg-level meaningful structure in the video data, and represents a high-level meaningful block in the video data.

[0020]

Referring now to FIG. 3, there is schematically illustrated the video signal processor according to the present invention. The video signal processor is generally indicated with a reference 10. In the video signal processor 10, the features of segments in the video data are used to determine the inter-segment similarity, group these segments into a scene, and automatically extract the video structure of the scene. Thus, the video signal processor 10 is applicable to both visual and audio segments.

[0021]

As shown in FIG. 3, the video signal processor 10 includes a video segmentor 11 to segment or divide an input video data stream into visual or audio segments or into both, a video segment memory 12 to store the segments of the video data, a visual feature extractor 13 to extract a feature for each visual segment, an audio feature extractor 14 to extract a feature for each audio segment, a segment feature memory 15 to store the features of the visual and audio segments, a scene detector 16 in which the visual and audio segments are grouped into a scene, and a feature similarity measurement block 17 to determine a similarity between two segments.

[0022]

The video segmentor 11 is supplied with a video data stream consisting of visual and audio data in any one of various digital formats including compressed video formats such as Moving Picture Experts Group Phase 1 (MPEG1), Moving Picture Experts Group Phase 2 (MPEG2) and digital video (DV), and divides the video data into visual or audio segments or into both segments. When the input video data is in

a compressed format, the video segmentor 11 can directly process the compressed video data without fully expanding it. The video segmentor 11 divides the input video data into visual or audio segments or into both segments. Also, the video segmentor 11 supplies the downstream video segment memory 12 with information segments resulted from the segmentation of the input video data. Further, the video segmentor 11 supplies the information segments selectively to the downstream visual feature extractor 13 and audio feature extractor 14, depending upon whether the information is visual or audio segments.

[0023]

The video segment memory 12 stores the information segments of video data supplied from the video segmentor 11. Also the video segment memory 12 supplies the information segments to the scene detector 16 upon query from the scene detector 16.

[0024]

The visual feature extractor 13 extracts a feature for each visual segment resulted from segmentation of the video data by the video segmentor 11. The visual feature extractor 13 can process a compressed video data without fully expanding it. It supplies the extracted feature of each visual segment to the downstream segment feature memory 15.

[0025]

The audio feature extractor 14 extracts a feature for each audio segment

resulted from segmentation of the video data by the video segmentor 11. The audio feature extractor 14 can process a compressed audio data without fully expanding it. It supplies the extracted feature of each audio segment to the downstream segment feature memory 15.

[0026]

The segment feature memory 15 stores the visual and audio segment features supplied from the visual and audio feature extractors 13 and 14, respectively. Upon query from the downstream feature similarity measurement block 17, the segment feature memory 15 supplies stored features value and segments to the feature similarity measurement block 17.

[0027]

The scene detector 16 groups the visual and audio segments into a scene using the information segments stored in the video segment memory 12 and the similarity between a pair of segments. Th scene detector 16 starts with each segment in a group to detect a repeated pattern of similar segments in a group of segments, and group such segments into the same scene. The scene detector 16 groups together segments into a certain scene, gradually enlarges the group until all the segments are grouped, and finally produces a detected scene for output. Using the feature similarity measurement block 17, the scene detector 16 determines how similar two segments are to each other.

[0028]

The feature similarity measurement block 17 determines a similarity between

two segments, and queries the segment feature memory 15 to retrieve the feature value for a certain segment.

[0029]

Since repeated similar segments lying close to each other in time are generally a part of the same scene, the video signal processor 10 detects such segments and groups them to detect a scene. The video signal processor 10 detects a scene by effecting a series of operations as shown in FIG. 4.

[0030]

First at step S1 in FIG. 4, the video signal processor 10 divides a video data into visual or audio segments as will be described below. The video signal processor 10 divides a video data supplied to the video segmentor 11 into visual or audio segments or possibly into both segments. The video segmenting method employed in the video signal processor 10 is not any special one. For example, the video signal processor 10 segments a video data by the method disclosed in the previously mentioned "G. Ahanger and T. D. C. Little: A Survey of Technologies for Parsing and Indexing Digital Video, Journal of Visual Communication and Image Representation 7: 28-4, 1996". This video segmenting method is well known in this field of art. The video signal processor 10 according to the present invention can employ any video segmenting method.

[0031]

Next at step S2, the video signal processor 10 extracts a feature. More

specifically, the video signal processor 10 calculates a pair of features for each segment to characterize the properties of a segment by means of the visual feature extractor 13 and audio feature extractor 14. The video signal processor 10 calculates, for example, a time duration of each segment, video or visual features such as color histogram and texture feature, audio features such as frequency analysis result, level and pitch, activity determination result, etc. as applicable features. Of course, the video signal processor 10 according to the present invention is not limited to these applicable features.

[0032]

Next at step S3, the video signal processor 10 measures a similarity between segments using their features. More specifically, the video signal processor 10 measures a dissimilarity between segments by the feature similarity measurement block 17 and determines how similar two segments are to each other according to the feature similarity measurement criterion of the feature similarity measurement block 17. Using the features having been extracted at step S2, the video signal processor 10 calculates a criterion for measurement of dissimilarity.

[0033]

At step S4, the video signal processor 10 groups the segments. More particularly, using the dissimilarity measurement criteria calculated at step S3 and features extracted at step S2, the video signal processor 10 iteratively groups similar segments lying close to each other in time. Thus, the video signal processor 10

provides a finally produced group as a detected scene.

[0034]

With the above series of operations, the video signal processor 10 can detect a scene from a video data. Therefore, using the above result, the user can sum the content of the video data and quickly access to points of interest in the video data.

[0035]

The operation of the video signal processor 10 at each of the steps shown in FIG. 4 will further be described below.

[0036]

First the video segmentation at step S1 will be discussed herebelow. The video signal processor 10 divides a video data supplied to the video segmentor 11 into visual or audio segments or into both segments if possible. Many techniques are available for automatic detection of a boundary between segments in a video data. As mentioned above, the video signal processor 10 according to the present invention is not limited to any special video segmenting method. On the other hand, the accuracy of scene detection in the video signal processor 10 substantially depends upon the accuracy of the video segmentation which is to be done before the scene detection. It should be noted that in the video signal processor 10, the scene detection can be done even with some error in the video segmentation. In this video signal processor 10, excessive segment detection which causes error is more preferable than insufficient one for the video segmentation. Namely, so long as excessive similar

segments are detected, even if the segment detection is not excessive, they can be grouped as those included in the same scene.

[0037]

Next the feature detection at step S2 will be discussed herebelow. The features are attributes of segments, characterizing the contents of the segments and providing information according to which a similarity between different segments is measured. In the video signal processor 10, the visual and audio feature extractors 13 and 14 calculate a pair of visual and audio features for each segment. However, the video signal processor 10 is not limited to any special features. The features considered to be effectively usable in the video signal processor 10 include visual feature, audio feature and visual- audio feature as will be described below. The requirement for these features usable in the video signal processor 10 is that they should be ones from which a dissimilarity can be determined. For a higher efficiency of signal processing, the video signal processor 10 simultaneously effects a feature extraction and video segmentation as the case may be. The features which will be described below meet the above requirement.

[0038]

The features include first a visual feature. Images (pictures) constituting a segment represent most of the content the segment depicts. Therefore, the similarity of the visual segment can often be converted to that of the image itself. Thus, the visual feature is an important one of the important features usable in the video signal

processor 10. The visual feature represents static information rather than dynamic information. Using a method which will be described later, the video signal processor 10 extracts the visual feature in the visual segment to obtain dynamic information.

[0039]

The visual features include many well-known ones. However, since it has been found that color feature (histogram) and video correlation, which will be described below, provide a good compromise between the cost and accuracy of calculation for the scene detection, the video signal processor 10 will use the color feature and video correlation as visual features.

[0040]

In the video signal processor 10, colors of images are important materials for determination of a similarity between two images. The use of a color histogram for determination of a similarity between images is well known as disclosed in, for example, "G. Ahanger and T. D. C. Little: A Survey of Technologies for Parsing and Indexing Digital Video, Journal of Visual Communication and Image Representation 7: 28-4, 1996". It should be noted that the color histogram is acquired by dividing a three-dimensional space such as HSV, RGB or the like for example into n areas and calculating a relative ratio of pixels of an image in each area. Information thus acquired gives an n-dimensional vector. Also, a color histogram can be extracted directly from a compressed video data as disclosed in the United States Patent No. 5,708,767.

[0041]

The video signal processor 10 samples, at a rate of 2 bits per color channel, an original YUV color space in images forming a segment to acquire a 64-length (= $2^{2\cdot3}$-length) histogram.

[0042]

Such a histogram represents a total color tone of an image but includes no spacial data. For this reason, a video correlation is calculated as one of visual features in the video signal processor 10. For the scene detection in the video signal processor 10, the interleaved segments provide an important index. For example, in a dialogue scene, the camera is moved between two persons alternately and to one of them being currently speaking. Usually, for shooting the same person again, the camera is moved back to nearly the same position where he or she was previously shot. Since it has been found that for detection of such a scene, a correlation of grayscale of images is a good index for a similarity between segments, the video signal processor 10 samples images to grayscale images each of M × N (both M and N may be small values; for example, M × N may be 8 × 8) in size to calculate a video correlation. The small gray scale images are interpreted as an MN-length feature.

[0043]

There is an audio feature, as the features different from the above-mentioned visual feature.

The audio feature can represent the content of an audio segment. In the video signal

processor 10, a frequency analysis, pitch, level, etc. is used as audio features. These audio features are known from various documents.

[0044]

First, the video signal processor 10 can make a frequency analysis of a Fourier Transform component or the like to determine the distribution of frequency information in a single audio frame. For example, the video signal processor 10 can use FFT (Fast Fourier Transform) component, frequency histogram, power spectrum and other features.

[0045]

Also, the video signal processor 10 may use pitches such as a mean pitch and maximum pitch, and levels such as mean loudness and maximum loudness, as effective audio features for representation of audio segments.

[0046]

Further features are those common to visual and audio segments. They are neither any visual feature nor audio feature, but provide useful information for representation of contents of segments included in a scene. The video signal processor 10 uses a segment length and an activity, as common visual-audio features.

[0047]

As in the above, the video signal processor 10 uses a segment length as a common visual-audio feature. The segment length is a time length of a segment. In the video signal processor 10, a scene has a rhythm feature of the change thereof, and

has the same tendency as that of the segment length in the scene. For example, short segments contiguous to each other with a short time between them represent a commercial program. On the other hand, segments included in a conversation or dialogue scene becomes longer, but the segments are similar to each other in the length. The video signal processor 10 can use such feature segment length as a common visual-audio feature.

[0048]

Also, the video signal processor 10 use an activity as a common visual-audio feature. The activity indicates how dynamic or static the content of a segment feels. For example, if a segment visually feels dynamic, the activity indicates a rapidity with which a camera is moved along an object or with which an object being shot by the camera changes.

[0049]

[0050]

[0051]

The activity is indirectly calculated by measuring a mean inter-frame dissimilarity in feature such as color histogram. A video activity V_F is given by the following equation (1):

$$V_F = \frac{\sum_{i=b}^{f-1} d_F(i,i+1)}{f-b} \qquad \text{.....} \qquad (1)$$

where i and j are frames, F is a feature measured between the frames i and j, $d_F(i, j)$ is a dissimilarity measurement criterion for the feature d_F, and b and f are numbers for a first frame and last frame in one segment. More specifically, the video signal processor 10 calculates the video activity V_F using the above-mentioned histogram for example.

[0052]

The features including the above-mentioned visual features basically indicate static information of a segment as in the above. To accurately represent the feature of a segment, however, dynamic information has to be indicated. For this reason, the video signal processor 10 represents dynamic information by a feature sampling method which will be described below.

[0053]

As shown in FIG. 5 for example, the video signal processor 10 extracts more than one pair of static features, starting at different time points in one segment. At this time, the video signal processor 10 determines the number of features to extract by keeping a balance between a highest fidelity and a minimum data redundancy. For example, when one image in the segment is designated as a key frame in that segment, a histogram calculated for the key frame is an extracted feature.

[0054]

Using the sampling method, of all values of the future, the video signal processor 10 determines which of the samples are to be selected in a segment. The the

video signal processor 10 needs a superior sampling method.

[0055]

Here, it will be considered that a certain sample is always taken at a predetermined time point, for example, at the last time point in a segment. In this case, samples from two arbitrary segments fading to black will be same black frames, so that no different features will possibly be acquired. That is, sampled two frames will be determined to be extremely similar to each other whatever the image contents of such segments are. This problem will take place since the samples are not good central values.

[0056]

For this reason, the video signal processor 10 is adapted not to extract a feature at such a fixed point but to extract a statistically central value. Here, the general feature sampling method will be described concerning two cases. That is, at first, a feature can be represented as a real-number n-dimensional vector, and secondly, only a dissimilarity measurement criterion can be used. It should be noted that best-known visual and audio features such as histogram, power spectrum, etc. are included in the features in the first case.

[0057]

In the first case, the number of samples is predetermined to be k and the video signal processor 10 automatically segments a feature of an entire segment into k different groups by using the well-known k-means clustering method as disclosed in

"L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: An Introduction to Cluster Analysis, John-Wiley and Sons, 1990". The video signal processor 10 selects one sample from each of the k groups. That is, the video signal processor 10 selects a centroid (mean vector) of the group or a sample near the centroid. The video signal processor 10 can perform the processing in a short time, and linear time is required in the number of samples.

[0058]

On the other hand, in the second case, the video signal processor 10 forms the k groups by the use of the k-medoids algorithm method also disclosed in "L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: An Introduction to Cluster Analysis, John-Wiley and Sons, 1990". The video signal processor 10 uses, as a sample value, the group medoid similar to the above-mentioned group centroid for each of the groups.

[0059]

It should be noted that in the video signal processor 10, the method for establishing the dissimilarity measurement criterion for extracted features is based on the dissimilarity measurement criterion for features on which the former method is based, which will further be described later.

[0060]

Thus, the video signal processor 10 can extract static features to represent dynamic information.

[0061]

As in the above, the video signal processor 10 can extract various features. However, each of such features is generally insufficient for representation, by itself, of a segment contents. For this reason, the video signal processor 10 can select a set of mutually complementary features by combining these different features. For example, the video signal processor 10 can provide more information than that of each feature by combining the above-mentioned color histogram and image correlation with each other.

[0062]

Next, the measurement of similarity between segments, in which the features acquired at step S3 in FIG. 4 are used, will be described herebelow. Using the dissimilarity measurement criterion being a function to calculate a real-number value with which it is determined how dissimilar two features are to each other, the video signal processor 10 measures a dissimilarity between the segments by means of the feature similarity measurement block 17. When the dissimilarity measurement criterion is small, it indicates that two features are similar to each other. If the criterion is large, it indicates that the two features are not similar to each other. The function for calculation of the dissimilarity between the two segments S_1 and S_2 concerning the feature F is defined as dissimilarity measurement criterion $d_F(S_1, S_2)$. This function has to meet the features given by the equations (2) below.

[0063]

$d_F(S_1, S_2) = 0$ (when $S_1 = S_2$)

$d_F(S_1, S_2) = 0$ (for all S_1 and S_2) ..... (2)

$d_F(S_1, S_2) = d_F(S_2, S_1)$ (for all S_1 and S_2)

[0064]

It should be noted that the appropriate dissimilarity measurement criteria sometimes depends on specific features. However, as disclosed in "G. Ahanger and T. D. C. Little: A Survey of Technologies for Parsing and Indexing Digital Video, Journal of Visual Communication and Image Representation 7: 28-4, 1996", and "L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: An Introduction to Cluster Analysis, John-Wiley and Sons, 1990", many general dissimilarity measurement criteria are effective in measuring a similarity between features represented as points in a n-dimensional space. The features include a Euclidean distance, inner product, L1 distance, etc. Since it is found that, of these features, the L1 distance will effectively act on various features including the histogram, image correlation, etc., the video signal processor 10 calculates the L1 distance $d_{L1}(A, B)$ between two n-dimensional vectors A and B by the following equation (3):

[0065]

$$d_{L1}(A,B) = \sum_{i=1}^{n} |A_i - B_i| \quad \text{..... (3)}$$

[0066]

where the subscript indicates the i-th element of each of the n-dimensional vectors.

[0067]

As mentioned above, the video signal processor 10 extracts, with respect to features which change as time elapse, features value at different time points in a segment. Then, to determine a similarity between two extracted features, the video signal processor 10 defines the dissimilarity of extracted features using a criterion for measurement of a dissimilarity of features on which the similarity measurement criterion is based. In many cases, the dissimilarity should most advantageously be established using a dissimilarity between a pair of features selected from each of the two extracted features and not most dissimilar to each other. That is, the video signal processor 10 determines the minimum dissimilarity. In this case, the sampled criterion for measurement of a dissimilarity between two extracted features SF_1 and SF_2 is given by the following equation (4):

[0068]

$$d_S(SF_1, SF_2) = \min_{F_1 \in SF_1, F_2 \in SF_2} d_F(F_1, F_2) \quad \text{..... (4)}$$

[0069]

The function $d_F(F_1, F_2)$ in the equation (4) above indicates a criterion for measurement of a dissimilarity between the extracted features F on which the equation (4) is based. It should be noted that the a maximum or mean value of the dissimilarity

may be taken instead of a minimum value as the case may be.

[0070]

In many cases, the video signal processor 10 needs to combine information derived from many features for the same segment to determine a similarity between segments. A solution for this problem is to calculate a weighting combination of various feature vectors dissimilarity function. That is, when there are available k features F_1, F_2, ..., F_k, the video signal processor 10 uses a dissimilarity measurement criterion $d_F(S_1, S_2)$ for combined features. The criterion is given by the following equation (5):

[0071]

$$d_F(S_1,S_2) = \sum_{i=1}^{k} w_i d_{Fi}(S_1,S_2) \quad \text{..... (5)}$$

[0072]

where $\{w_i\}$ is a pair of weigh and $\Sigma_i w_i=1$.

[0073]

As in the above, the video signal processor 10 can calculate a dissimilarity measurement criterion using features having been extracted at step S2 in FIG. 4 to determine a similarity between segments in consideration.

[0074]

Next, the segment grouping at step S4 in FIG. 4 will be described herebelow. Using the dissimilarity measurement criterion and extracted features, the video signal

processor 10 repeatedly combines similar segments lying close to each other in time, groups these segments, and outputs a finally produced group as a detected scene.

[0075]

When detecting a scene by grouping segments, the video signal processor 10 effects two basic operations. One of the operations is to detect groups of similar segments lying close to each other in time. Most of the groups thus acquired will be a part of the same scene. The other operation effected in the video signal processor 10 is to combine concurrent and similar scenes together since the segments are concurrent. The video signal processor 10 starts these operations from each segment, and repeats them. Then the video signal processor 10 organizes a step-by-step larger group of segments and outputs a finally produced group as a set of scenes.

[0076]

To control these operations, the video signal processor 10 is controlled under the following two constraints.

[0077]

Under one of the two constraints, the video signal processor 10 has to adopt a dissimilarity threshold δ_{sim} to determine whether two similar segments are thought to be similar enough to belong to the same scene. As shown in FIG. 6 for example, the video signal processor 10 judges whether one of the segments is similar or not similar to the other.

[0078]

It should be noted that the video signal processor 10 may be adapted to set the dissimilarity threshold δ_{sim} by the user or automatically as will be described later.

[0079]

Under the second constraint, the video signal processor 10 has to adopt a temporal threshold T based on which it is determined that the two segments separated apart can be considered to be included in the same scene. As shown in FIG. 7 for example, the video signal processor 10 puts, into the same scene, two similar segments A and B lying close to each other within the temporal threshold T but not two segments B and C similar to each other but having between them a time gap not within the temporal threshold T. Thus, because of the constraint by the temporal threshold T, the video signal processor 10 will not erroneously put into the same scene two segments similar to each other but largely apart from each other on the time base.

[0080]

Since it has been found that 6 to 8 set as the temporal threshold T would generally give a good result, the video signal processor 10 uses the temporal threshold T for 6 to 8 in principle.

[0081]

It is assumed herein that to acquire a group of similar segments, the video signal processor 10 adopts the hierarchical clustering method disclosed in "L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: An Introduction to Cluster Analysis, John-Wiley and Sons, 1990". In this algorithm, a criterion $d_c(C_1, C_2)$ for determination of

a dissimilarity between two clusters C_1, C_2 is defined as a minimum similarity between included elements. It is given by the following equation (6):

[0082]

$$d_c(C_1,C_2) = \min_{S_1 \in C_1, S_2 \in C_2} dist_S(S_1,S_2) \quad \text{..... (6)}$$

[0083]

It should be noted that in the video signal processor 10, a minimum function expressed by the equation (6) can easily be replaced with a maximum function or mean function.

[0084]

First at step S11 in FIG. 8, the video signal processor 10 initializes a variable N to the number of segments. The variable N indicates the concurrent number of groups always detected.

[0085]

Next at step S12, the video signal processor 10 generates a set of clusters. In the initial state, the video signal processor 10 takes N segments as different from each other. That is, there exist N clusters in the initial state. Each of the clusters has features indicating a start time and end time represented by C^{start} and C^{end}, respectively. Elements included in the set of clusters are arranged in order starting from the C^{start}.

[0086]

Next at step S13, the video signal processor 10 initializes a variable t to 1. At step S14, the video signal processor 10 judges whether the variable t is larger than the temporal threshold T. If the video signal processor 10 determines that the variable t is larger than the temporal threshold T, it will go to step S23. When it determines that the variable t is smaller than the temporal threshold T, it will go to step S15. Since the variable t is 1, however, the video signal processor 10 will go to step S15.

[0087]

At step S15, the video signal processor 10 calculates the dissimilarity measurement criterion d_c to detect two of the N clusters that are the most similar to each other. Since the variable t is 1, however, the video signal processor 10 will calculate the dissimilarity measurement criterion d_c between adjacent clusters to detect among the adjacent clusters a pair of clusters that are the most similar to each other.

[0088]

An approach to detect two clusters which are the most similar to each other may be to scan all possible pairs of object clusters. Having the constraint by the temporal threshold T, the video signal processor 10 can restrict the number of pairs of the object clusters. So, the video signal processor 10 has only to scan clusters which are separated by t segments. Since the pairs of clusters are arranged in the temporal order, the video signal processor 10 scans forward and backward for a certain cluster to scan clusters separated more than t segments. Thus the entire segments after these clusters

will be outside the object and scanning will be completed.

[0089]

The two clusters thus detected are defined as C_i and C_j, respectively, and a dissimilarity between the clusters C_i and C_j is defined as d_{ij}.

[0090]

At step S16, the video signal processor 10 will judge whether the dissimilarity d_{ij} is larger than the dissimilarity threshold δ_{sim}. When the dissimilarity d_{ij} is judged larger than the dissimilarity threshold δ_{sim}, the video signal processor 10 will go to step S21. If the the dissimilarity d_{ij} is judged smaller than the dissimilarity threshold δ_{sim}, the video signal processor 10 will go to step S17. It is assumed here that the dissimilarity d_{ij} is smaller than the dissimilarity threshold δ_{sim}.

[0091]

At step S17, the video signal processor 10 will merge the cluster C_j into the cluster C_i. That is, the video signal processor 10 will add to the cluster C_i all the elements in the cluster C_j.

[0092]

Next at step S18, the video signal processor 10 will remove the cluster C_j from the set of clusters. It should be noted that if the start time C_i^{start} changes due to the combination of the two clusters C_i and C_j, the video signal processor 10 will rearrange the elements in the set of clusters to keep the order at the start time.

[0093]

Next at step S19, the video signal processor 10 will subtract 1 from the variable N.

[0094]

At step S20, the video signal processor 10 will judge whether the variable N is 1 or not. If the variable N is judged to be 1, the video signal processor 10 will go to step S23. When the video signal processor 10 determines that the variable N is not 1, it will go to step S15. It is assumed here that the variable N is not 1.

[0095]

Thus, at step S15, the video signal processor 10 will calculate the dissimilarity measurement criterion d_c again to detect two clusters the most similar to each other. Since the variable t is 1, the video signal processor 10 will calculate the criterion d_c for determination of the dissimilarity between adjacent clusters to detect a pair of clusters that are the most similar to each other.

[0096]

Next at step S16, the video signal processor 10 will judge whether the dissimilarity d_{ij} is larger than the dissimilarity threshold δ_{sim}. It is also assumed here that the dissimilarity d_{ij} is smaller than the dissimilarity threshold δ_{sim}.

[0097]

The video signal processor 10 will effect the operations at steps S17 to S20.

[0098]

When as a result of the repetition of the above operations and subtraction of 1

from the variable N, it is determined at step S20 that the variable N is 1, the video signal processor 10 will go to step S23 where it will combine together clusters each including a single segment. That is, in this case, since all segments are grouped into one cluster, the video signal processor 10 does not have to perform the operation and the series of operations are terminated.

[0099]

If the video signal processor 10 determines at step S16 that the dissimilarity d_{ij} is larger than the dissimilarity threshold δ_{sim}, it will go to step S21 where it will repeatedly combine clusters which concurrently exist. Namely, if the time interval between C_i^{start} and C_i^{end} of the cluster C_i is concurrent with that between C_j^{start} and C_j^{end} of the cluster C_j, since the two clusters C_i and C_j overlap each other, the video signal processor 10 can detect concurrent clusters and combine the clusters together by arranging the clusters in a set based on the start time of the cluster set.

[0100]

At step S22, the video signal processor 10 will add 1 to the variable t which will thus be t = 2, and go to step S14 where it will judge whether the variable t is larger than the temporal threshold T. It is also assumed here that the variable t is smaller than the temporal threshold T and the video signal processor 10 will go to step S15.

[0101]

At step S15, the video signal processor 10 will calculate the dissimilarity measurement criterion d_c and detects two of a plurality of clusters existing currently,

that are the most similar to each other. However, since the variable t is 2, the video signal processor 10 calculates the criterion d_c for determination of the dissimilarity between every other clusters to detect a pair of clusters the most similar to each other.

[0102]

Then at step S16, the video signal processor 10 judges whether the dissimilarity d_{ij} between every other clusters C_i and C_j is larger than the dissimilarity threshold δ_{sim}. It is assumed here that the dissimilarity d_{ij} is smaller than the dissimilarity threshold δ_{sim}. After effecting the operations at steps S21 and S22, the video signal processor 10 adds 1 to the variable t which will thus be $t = 3$, and will move to step S14 and subsequent steps. It is assumed that, when the variable t is 3, the video signal processor 10 will calculate, at step S15, the criterion d_c for determination of the dissimilarity between every two clusters, and detect a pair of clusters which are the most similar to each other.

[0103]

When as a result of the repetition of the above operations and addition of 1 to the variable t, it is determined at step S14 that the variable t is larger than the tine threshold T, the video signal processor 10 will go to step S23 where it will combine clusters each including a single segment. That is, the video signal processor 10 will take discrete clusters as ones each including a single element. If there exist a sequence of such clusters, the video signal processor 10 will combine them together into a single ones. This process combines together segments having no relation in similarity with

any single scene. However, it should be noted that the video signal processor 10 has not to always effect this process.

[0104]

The video signal processor 10 terminates the series of operations by way of step S22.

[0105]

With this series of operations, the video signal processor 10 can gather the plurality of clusters and generate a scene to be detected.

[0106]

It should be noted that the video signal processor 10 may be adapted to set the dissimilarity threshold δ_{sim} by the user or automatically determine it as having previously been described. However, when a fixed value is used, the optimum value of the dissimilarity threshold δ_{sim} will depend upon the content of a video data. For example, for a program whose content is variable, the dissimilarity threshold δsim has to be set to a high value. On the other hand, for a program having a less-variable content, the dissimilarity threshold δ_{sim} has to be set to a lower value. Generally, when the dissimilarity threshold δ_{sim} is too high, scenes to be detected will be too little. On the other hand, when dissimilarity threshold δ_{sim} is too low, scenes to be detected will be too much.

[0107]

Thus, an optimum dissimilarity threshold δ_{sim} has to be determined since the

performance of the video signal processor 10 depends greatly upon the dissimilarity threshold δ_{sim}. Therefore, when the video signal processor 10 is adapted to set a dissimilarity threshold δ_{sim} by the user, the above has to be taken in consideration. On the other hand, the video signal processor 10 may be adapted to automatically set an effective dissimilarity threshold δ_{sim} by using any of methods which will be described below.

[0108]

One of the methods will be described by way of example. Namely, the video signal processor 10 acquire a dissimilarity threshold δ_{sim} from the distance distribution expressing the difference of the similarity between $(n)(n-1)/2$ segment pairs by using statical future such as mean value and mode. For example, the video signal processor 10 sets the dissimilarity threshold δ_{sim} to $a\mu + b\sigma$ by using mean value μ and standard deviation σ of the distance expressing the difference of the similarity between all segment pairs, where a and b are fixed constants. It has been found that setting of the constants a and b to 0.5 and 0.1, respectively, will assure a good result.

[0109]

In practice, the video signal processor 10 has not to determine the distance expressing the difference of the similarity between all pairs of segments. Instead of all pairs of segments , the video signal processor 10 extracts subset of distance value for judging the real mean value μ and standard deviation σ at random. That

is, when calculating the distance, the video signal processor 10 selects two segments at random. In case the video signal processor 10 selects two segments sufficiently, nearly real mean value μ and standard deviation σ can be acquired and an appropriate dissimilarity threshold δ_{sim} can be automatically acquired. In this case, the video signal processor 10 sets the total number of segments to n, an arbitrary small constant to C, and can automatically determine an appropriate dissimilarity threshold δ_{sim}.by extracting Cn pairs of segments.

[0110]

In the foregoing, the use of a single dissimilarity measurement criterion in the video signal processor 10 has been described. In addition, the video signal processor 10 can a weighting function to combine a variety of dissimilarity measurement criteria for different types of features in order to judge whether segments in pairs are in the same group, as having previously been described. The features can only be weighted after trial and error, and when the features are different in type from each other, it is usually difficult to appropriately weight them. However, using a color histogram and texture feature, for example, in combination, the video signal processor 10 can detect possible scenes for these features and combine a single scene structure from the structures of the detected scenes. The results of scene detection for the features will be referred to as "scene layer" hereinafter.

For example, when a color histogram and segment length are used as features, the video signal processor 10 can detect scenes for these features to provide a scene layer for the color histogram and a one for the segment length to combine these scene layers into a single scene structure.

[0111]

Generally, information from video and audio domains cannot be combined in principle. Using a similar method to that for combining features different in quality from each other, the video signal processor 10 can combine into a single scene structure scene layers obtainable from video and audio domains.

[0112]

Such a processing will be described herebelow. It is assumed here that there are k features F_1, F_2, ..., F_k each representing one similarity criterion and there are available a dissimilarity measurement criterion d_F^i, dissimilarity threshold δ^i_{sim} and a temporal threshold T^i correspondingly to the features F_i. Using the dissimilarity measurement criterion d_F^i, dissimilarity threshold δ^i_{sim} and a temporal threshold T^i for the features F_i, the video signal processor 10 detects a set of scene layers $X_i = \{X_i^j\}$. It is assumed that the video signal processor 10 detects divisional scene layers for video and audio information, respectively, and generates two independent scene layers $X_i = \{X_i^j\}$ (i = 1, 2) for the video and audio information, respectively.

[0113]

The video signal processor 10 has to determine how to combine scene

boundaries for combination of difference scene layers into a single scene structure. The scene boundaries do not always match one another. It is assumed here that for the scene layers, there exists boundary points t_{i1}, t_{i2}, ..., $t_i|x_i|$ represented by a sequence of times indicating the scene boundaries. The video signal processor 10 first selects a certain scene layer to be a basis for alignment in order to combine various scene layers into a single group. Then, the video signal processor 10 determines for each of the boundary points t_{i1}, t_{i2}, ..., $t_i|x_i|$ whether they are those in the scene structure produced by finally combining the scene layers.

[0114]

It is assumed here that, for each i as the number of the set of scene layers, the Boolean function indicating whether the i-th scene layer X_i has a scene boundary near a time t is $B_i(t)$. The term "near" varies and it is for example 0.5 sec when video and audio information are matched, respectively.

[0115]

The video signal processor 10 calculates the logical function $B_1(t_j)$ for each of the boundary points $t_j = t_{ij}$ when $j = 1, ..., |X_j|$ and $i = 1, ..., k$. The calculation result will indicate whether the scene boundary exists near the time t in the scene layer X_i for each of the independent scene layers. The video signal processor 10 uses, as a decision function, $B_i(t_j)$ to determine whether, in the combined decision, the time t is a scene boundary.

[0116]

A simple example of the decision function is to calculate the sum of $B_i(t_j)$ which is more than a constant m and equal to 1. The time point t designates scene boundary in the last scene structure. Especially when the m = 1, it means that the boundary point is equal to "or-ing" boundary point. On the other hand, when m = k, it means that all the boundary points are required to be made even.

[0117]

Thus, the video signal processor 10 can combine difference scene layers into a single scene structure.

[0118]

As having been described in the foregoing, the video signal processor 10 according to the present invention is to extract a scene structure, and can be applied to video data contents of different types. Experiments using actual video data contents has been carried out, and it has already been proved that a scene structure can be restored from other data genre such as TV dramas, movies, etc. by the video signal processor 10.

[0119]

The video signal processor 10 is full automatic and can automatically determine an appropriate threshold correspondingly to a change in content of a video data without the necessity of any user's intervention to set the aforementioned dissimilarity threshold and temporal threshold.

[0120]

Further, the video signal processor 10 according to the present invention can be operated by the user without any prior knowledge of structure of video data contents.

[0121]

Moreover, since the video signal processor 10 is very simple and efficient in calculation, so it can be applied in home electronic appliances such as a set-top box, digital video recorder, home server, etc.

[0122]

Also the video signal processor 10 can provide a result of scene detection as a basis for a new high-level access for the video browsing. Therefore, the video signal processor 10 permits an easy access to a video data, which is based on the content of the data, by imaging the content of the video data using the high-level scene video structure, not any segments. For example, the video signal processor 10 displays a scene by which the user can quickly know the summary of a program and thus quickly find a part of the program in which he is interested.

[0123]

Furthermore, the video signal processor 10 can provide a result of scene detection as a basis for automatic outlining or digesting of a video data. Different from combining random fragments from video data, for a consistent summing-up, it is generally necessary to decompose a video data into reconstructible meaningful components. A scene detected by the video signal processor 10 serves as a normal

basis for preparation of such a digest.

[0124]

It should be noted that the present invention is not limited to the embodiment having been described in the foregoing, but the features used for measurement of similarity between segments example may of course be other than those having been described and be appropriately modified without departing from the scope of the present invention defined later.

[0125]

[Effect of the Invention]

As having been described in detail in the foregoing, the present invention provides the signal processing method for detecting and analyzing a pattern reflecting the semantics of the content of a signal, the method including steps of extracting, from a segment consisting of a sequence of consecutive frames forming together the signal, at least one feature information which characterizes the properties of the segment; calculating, using the extracted feature information, a criterion for measurement of a similarity between a pair of segments for every extracted feature information and measuring a similarity between a pair of segments according to the similarity measurement criterion; and detecting, according to the feature information and similarity measurement criterion, two of the segments, whose mutual time gap is within a predetermined temporal threshold and mutual similarity is not less than a predetermined dissimilarity threshold, and

grouping the segments into a scene consisting of a sequence of temporally consecutive segments reflecting the semantics of the signal content.

[0126]

Therefore, the signal processing method according to the present invention can detect similar segments in a signal and group them into a scene, thereby permitting to extract a higher-level structure than a segment.

[0127]

Also the present invention provides the video signal processor for detecting and analyzing a visual and/or audio pattern reflecting the semantics of the content of a supplied video signal, the apparatus including means for extracting, from a visual and/or audio segment consisting of a sequence of consecutive visual and/or audio frames forming together the video signal, at least one feature information which characterizes the properties of the visual and/or audio segment; means for calculating, using the extracted feature information, a criterion for measurement of a similarity between a pair of visual segments and/or audio segments for every extracted feature information and measuring a similarity between a pair of visual segments and/or audio segments according to the similarity measurement criterion; and means for detecting, according to the feature information and similarity measurement criterion, two of the visual segments and/or audio segments, whose mutual time gap is within a predetermined temporal threshold and mutual similarity is not less than a predetermined dissimilarity threshold, and grouping the visual

segments and/or audio segments into a scene consisting of a sequence of temporally consecutive visual segments and/or audio segments reflecting the semantics of the video signal content.

[0128]

Therefore, the video signal processor according to the present invention can detect similar visual segments and/or audio segments in the video signal and group them for output as a scene, thereby permitting to extract a higher-level video structure than a visual and/or audio segment.

[Brief Description of the Drawing]

FIG. 1 explains the structure of a video data to which the present invention is applicable, using a video data model.

FIG. 2 explains a scene.

FIG. 3 is a block diagram of an embodiment of the video signal processor according to the present invention.

FIG. 4 is a flow chart of a series of operations effected in detecting segments and grouping them into a scene in the video signal processor.

FIG. 5 explains the sampling of dynamic features in the video signal processor.

FIG. 6 explains the dissimilarity threshold.

FIG. 7 explains the temporal threshold.

FIG. 8 is a flow chart of a series of operations effected in grouping segments

in the video signal processor.

[Description of the Numerals]

10 video signal processor

11 video segmentor

12 video segment memory

13 visual feature extractor

14 audio feature extractor

15 segment feature memory

16 scene detector

17 feature similarity measurement block

[Name of document] Abstract

[Summary]

[Task]

To extract a high-level video structure in a variety of video data.

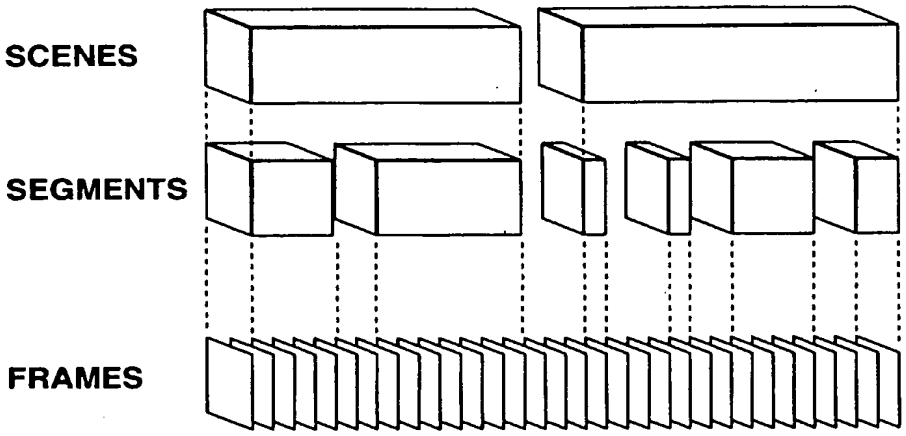
[Solving Means]

The video signal processor 10 includes a scene detector 16 which uses features extracted for visual segments and/or audio segments resulted from segmentation of an input stream of video data, and a criterion for measurement of similarity between visual and/or audio segment pairs, calculated for each of the features using the similarity measurement criterion, to detect two visual segments and/or audio segments whose time gap is within a predetermined temporal threshold and whose similarity is not less than a predetermined dissimilarity threshold and group the segments into a scene consisting of visual segments and/or audio segments reflecting the semantics of the video data content and temporally contiguous to each other.

[Selecting Drawing] Fig.3

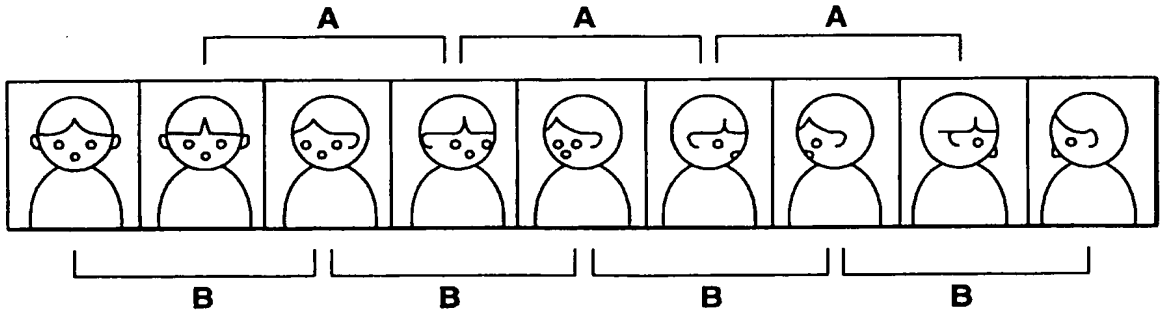
[DOCUMENT NAME] DRAWING

[FIG. 1]



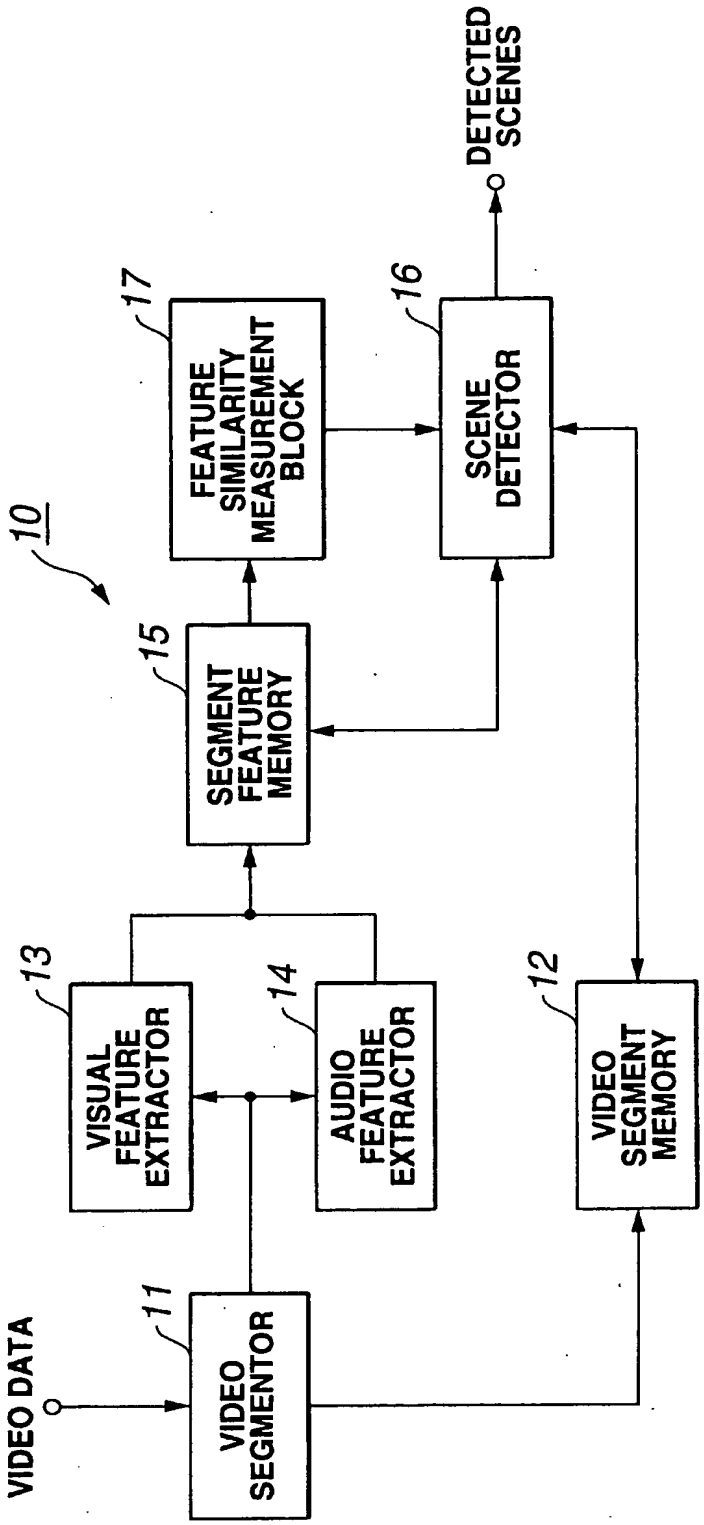
HIERARCHY MODEL OF THE VIDEO STRUCTURE

[FIG. 2]



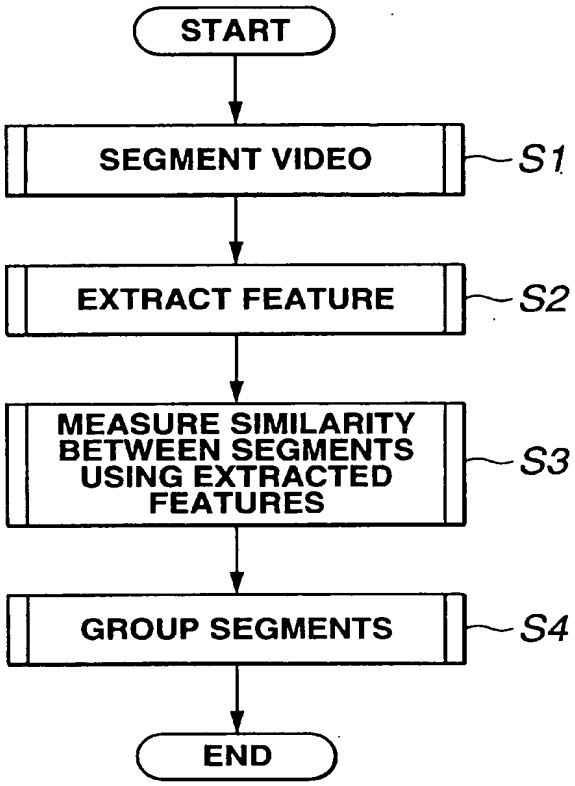
EXPLANATION DRAWING OF A SCENE

[FIG. 3]



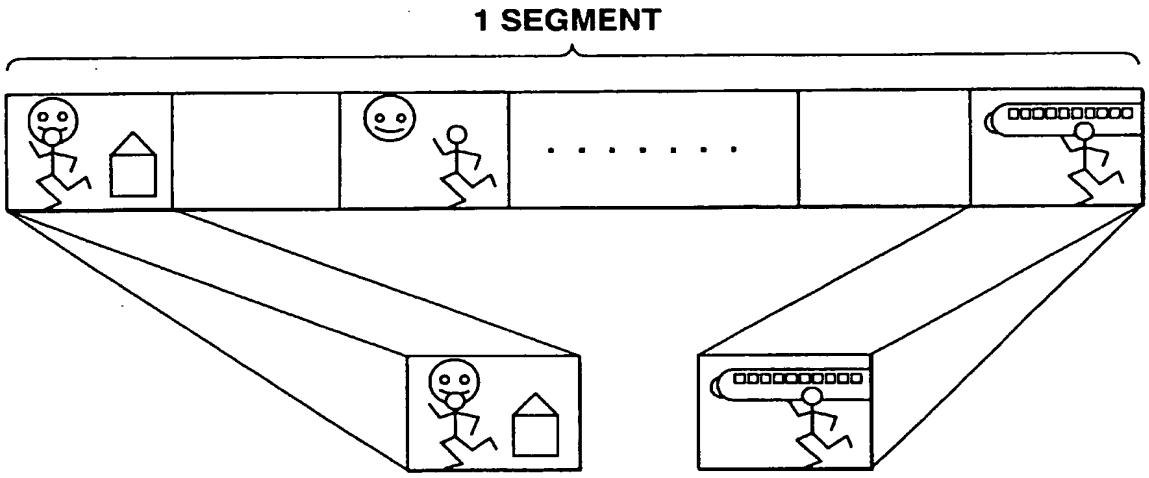
BLOCK DIAGRAM OF THE VIDEO SIGNAL PROCESSOR

[FIG. 4]



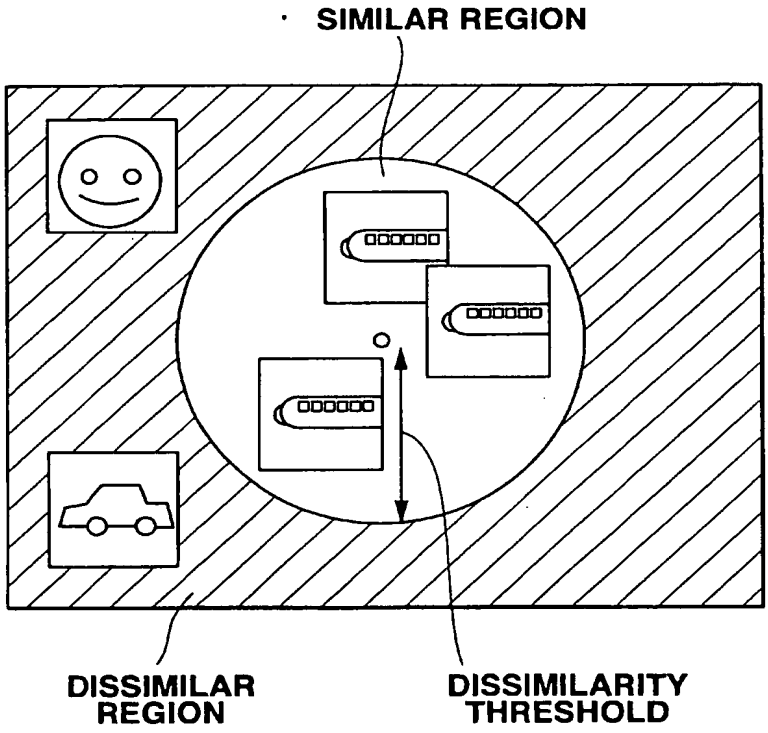
A SERIES OF OPERATIONS IN THE VIDEO SIGNAL PROCESSOR

[FIG. 5]



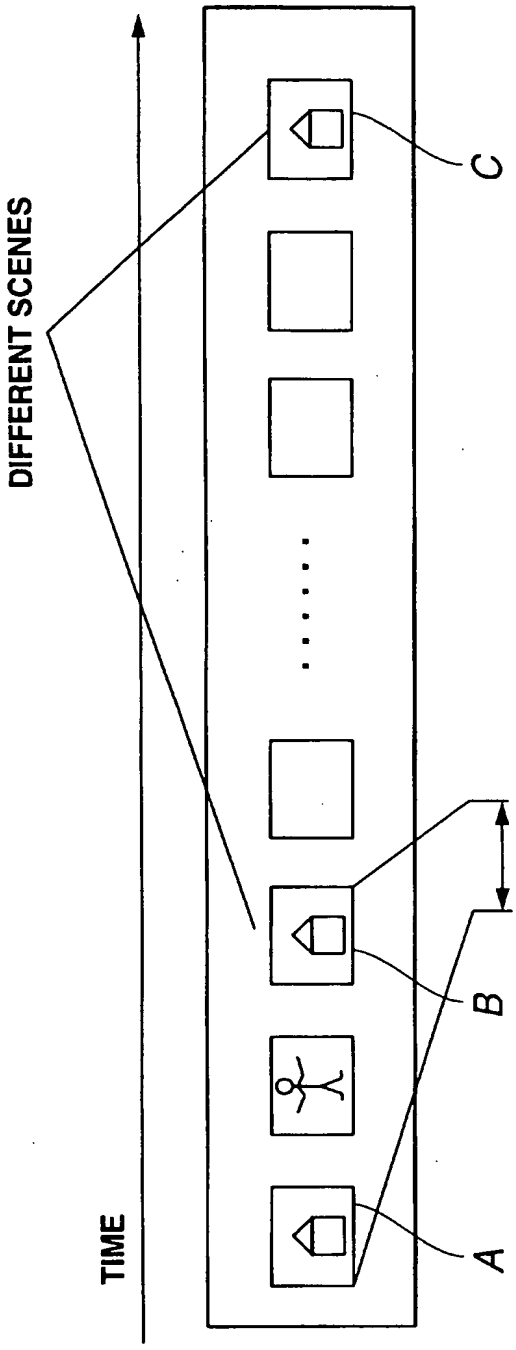
EXPLANATION DRAWING OF SAMPLING METHOD OF FEATURES

[FIG. 6]



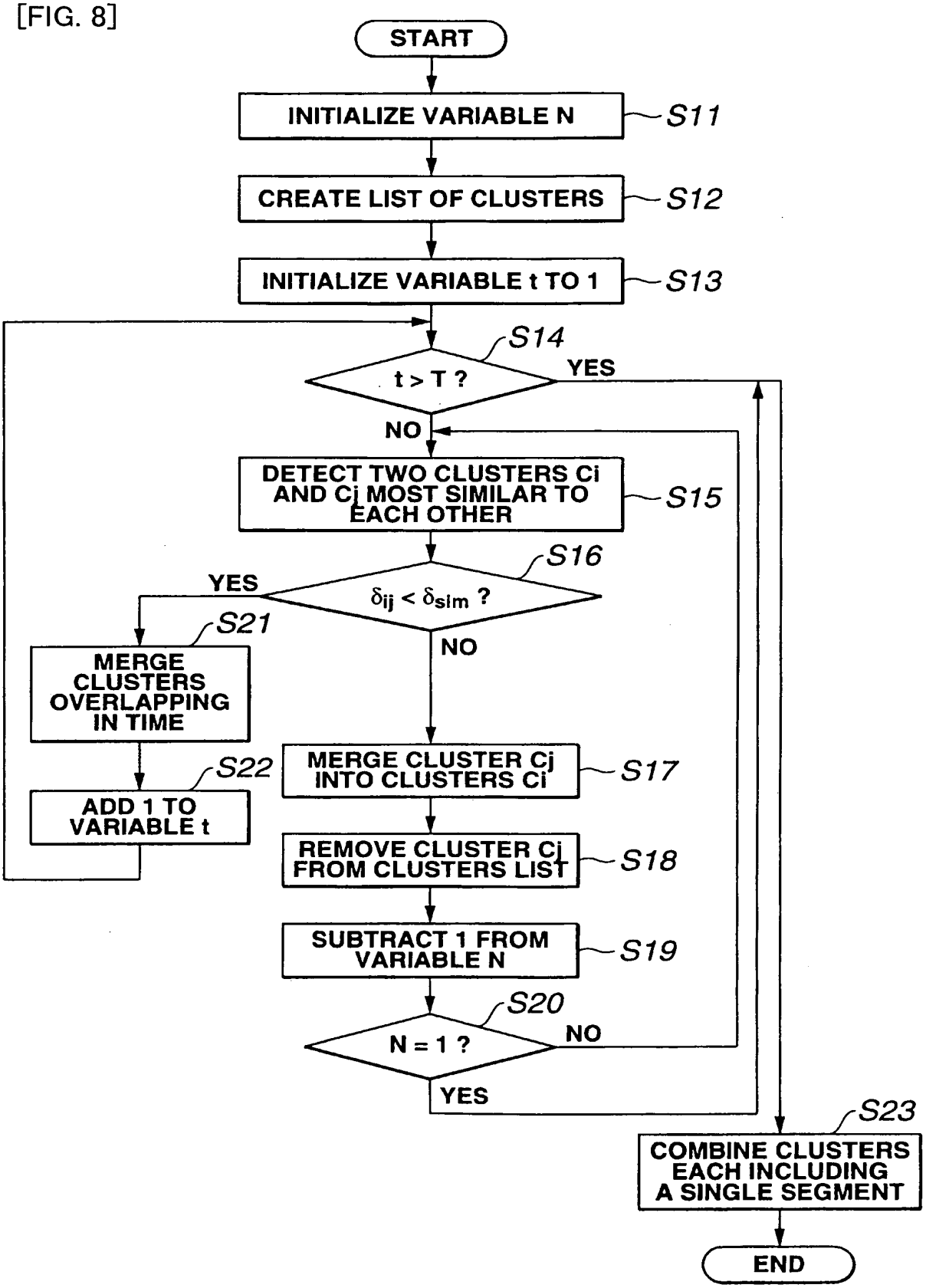
EXPLANATION DRAWING OF THE DISSIMILARITY THRESHOLD

[FIG. 7]



EXPLANATION DRAWING OF THE TEMPORAL THRESHOLD

[FIG. 8]



A SERIES OF OPERATIONS IN THE VIDEO SIGNAL PROCESSOR

Information of Record for Applicant

Identification Number: [000002185]

1. Date of Change: August 30, 1990

[Reason of Change] Registration

[Address] 7-35,Kitashinagawa 6-chome,

Shinagawa-ku, Tokyo, Japan

[Name] Sony Corporation